Beyond LLM

知未智能 KDF 2024/04

让GPT告诉我：如何超越LLM

You
从应用创新的角度来看，如何超越LLM

ChatGPT
从应用创新的角度来看，要想超越大型语言模型（LLM），需要关注几个关键的方面。这包括提高模型的适应性、多样性、可解释性以及用户交互的质量。以下是一些可能的策略：

1. 定制化和适应性

- **域特定模型**：开发专门针对特定领域或行业的模型，以提供更准确和相关的输出。
- **个性化学习**：通过观察用户与模型的互动来定制化模型的响应，使其更好地适应个别用户的需求和偏好。

2. 增强多样性和创新

- **跨模态能力**：集成视觉、听觉和其他感知模式的处理能力，以实现更丰富的用户体验和应用场景。
- **生成创新内容**：利用LLM在艺术、设计和其他创造性领域生成原创作品，鼓励创新。

3. 提升可解释性和可信度

- **可解释的AI**：开发技术来解释模型的决策过程，增加用户对模型输出的信任。
- **事实验证和误差校正**：集成事实验证组件来识别和更正误导性或不准确的信息，提高输出质量。

4. 改进用户交互

- **自然语言理解（NLU）的提升**：改进模型的理解能力，使其更好地捕捉用户的意图和上下文。
- **多轮对话管理**：开发高级对话管理系统，使模型能够更自然和连贯地参与到长期对话中。

5. 遵守伦理和社会责任

- **伦理和透明度**：确保模型的开发和部署遵循伦理准则，强化对用户隐私和数据安全的保护。
- **减少偏见和不公平**：积极识别和减少模型偏见，促进公平性和多样性。

6. 技术创新

- **优化算法和架构**：通过研究新的算法和网络架构来提高模型效率和减少资源消耗。
- **增强学习和反馈循环**：利用增强学习和用户反馈循环不断优化模型性能。

7. 互动式学习

- **即时反馈调整**：允许用户对模型的回答提供即时反馈，模型根据反馈实时调整，以提供更准确和个性化的信息。
- **协同学习**：通过用户与模型之间的互动学习，让模型更好地理解人类的语言习惯和沟通方式。

8. 知识增强

- **动态知识库集成**：将模型与实时更新的知识库连接，使其能够提供最新的信息和数据。
- **上下文感知能力**：发展模型的能力以理解和利用用户的上下文环境，包括地理位置、时间、以往的交互历史等，以提供更相关的回应和建议。

9. 多语言和跨文化能力

- **多语言支持**：提高模型的多语言处理能力，使其能够理解和生成多种语言的内容，扩大其应用范围。
- **跨文化理解**：使模型能够理解和尊重不同文化的差异和细微之处，提高跨文化交流的准确性和敏感性。

10. 高级情感计算

- **情绪识别和响应**：发展模型的能力以识别用户的情绪状态，并以适当的情绪反应回应，以提高人机交互的自然度和舒适度。
- **情绪智能**：在生成的内容中嵌入适当的情感，使得机器生成的文本、音频或视频更加生动和人性化。

11. 安全和隐私保护

- **强化安全性**：通过加强数据加密和匿名处理，保护用户数据免受未经授权的访问和滥用。
- **隐私保护机制**：开发和实施高级的隐私保护技术和政策，确保用户的隐私不被侵犯。

12. 可扩展性和灵活性

- **云端和边缘计算**：利用云计算和边缘计算的优势，提高模型的可扩展性和响应速度，同时降低延迟。
- **模块化设计**：通过模块化设计使系统易于更新和维护，支持新功能的快速集成。

13. 生态系统和平台建设

- **开放API和平台**：通过提供开放的API和开发平台，鼓励外部开发者和企业开发和共享新的应用和服务。
- **生态系统合作**：与其他技术供应商、研究机构和教育机构建立合作伙伴关系，共同推动AI技术的发展和应用。

14. 集成先进的推理能力

- **逻辑推理**：加强模型在复杂逻辑和数学问题上的推理能力，使其能够进行更高级的推理和解题。
- **常识推理**：增强模型的常识推理能力，让它更好地理解和处理日常生活中的问题和场景。

15. 专注于持续学习

- **在线学习**：开发模型能力以实时学习新信息和技能，而不需要进行昂贵的重新训练或更新。
- **终身学习系统**：构建模型能够持续学习和适应新任务和环境变化的能力，降低长期维护成本。

16. 促进人机协作

- **增强型合作**：开发使人类与AI能够更紧密合作的工具和接口，如协助决策、共同创作和问题解决。
- **角色适应与分工**：让AI系统能够根据场景和需求自动调整其角色和贡献，优化人机协作效率。

17. 发展元学习能力

- **学习学习**：让模型具备元学习能力，即学习如何更有效地学习新任务，从而加速技能的获取和适应新领域。
- **跨任务迁移学习**：通过跨任务学习和迁移学习提升模型的灵活性，使其能够在完成一个任务的同时，快速适应另一个不同的任务。

18. 优化资源效率

- **算法和硬件协同优化**：在算法层面和硬件设计层面进行协同优化，以提高计算效率和降低能耗。
- **稀疏模型和压缩技术**：利用稀疏性技术和模型压缩方法减少模型的大小和运行成本，使AI更加可持续和普及。

19. 探索新型交互方式

- **增强现实与虚拟现实**：集成AR和VR技术，创造沉浸式交互体验，为教育、训练和娱乐等领域开辟新的可能性。
- **语音和手势控制**：开发更自然的交互界面，如高级语音识别和手势控制，以提升用户体验和接入便利性。

20. 推动道德和社会责任

- **道德设计**：在AI系统的设计和开发过程中主动考虑和实施道德原则和社会价值观。
- **社会影响评估**：定期进行社会影响评估，确保技术发展不仅符合经济目标，同时也促进社会福祉和公平正义。

https://chat.openai.com/share/1501ecb4-1b73-43f1-965d-ec3669343495

LLM的问题

因为没有上下文信息，无法给出更精确的结果

给出了指导方向，无法直接给出解决方案

跨模态能力

集成视觉、听觉和其他感知模式的处理能力，以实现更丰富的用户体验和应用场景。

发言人A:

根据路透报道说，这个美国的这个发言人，Herbie，他说。有报道指出，以色列军方都在几乎没有人监督的情况之下。在用AI帮助识别可疑的极端分子跟目标，所以，有数万加薩人被标记为轰炸的目标。然后，他就说，美国现在正在展开调查。我先要请教一下赖老师，因为这个人道援助的这个队伍才刚被攻击。然后就说，其实这是蓄意所为的吧，引起很大的争论。结果，那你看，他们是不是真的，其实根本都早已是用AI在锁定的目标?

这家公司的发展历程是什么样

详细介绍一下这家公司

宁波兴瑞电子科技股份有限公司成立于2001年，前身为1990年创立的慈溪市高王中兴电子器材厂。公司深耕于精密制造业长达三十余年，是一家具备一站式全制程研发、设计和制造能力的企业，为客户提供高度定制化的产品和服务。公司以模具开发制造为技术核心，采用精密注塑/冲压和自动化组装等先进技术，聚焦新能源汽车电装系统、智能终端、消费电子等赛道，为下游客户提供电子连接器、结构件、镶嵌注塑件等，广泛覆盖汽车电子和智能终端等领域。

公司产品覆盖多领域，包括汽车电装领域和智能终端领域。在汽车电装领域，公司提供的产品主要包括连接器、液晶框架、IBMU/BDU/DCDC等镶嵌注塑零组件、电控周边零组件及传感器组件等，服务的客户涵盖了Panasonic、HELLA、BOSCH、TADITEL等国际汽车电子龙头公司以及中车时代等国内Tier 1汽车电子客户。智能终端领域的主要产品包括智能机顶盒、网通网关、智能安防、智能电表等，下游客户涵盖了仁宝电脑、特艺（Vantiva）、萨基姆（Sagemcom）、松下（Panasonic）等智能终端领域的头部厂商。

公司深度绑定宝马、通用凯迪拉克等海外头部车企，并在为头部客户提供高附加值的嵌塑集成产品方面取得突破。同时，公司也在加快拓展国内汽车客户，成功打入国内新能源汽车电子供应链，现已陆续获得中车时代及国内电装头部企业的项目定点，终端客户包括小米汽车、比亚迪等国内厂商。

2018年，公司登陆深交所上市，正式进入国内资本市场。在产能布局中，公司放眼全球，始终稳步践行国际化、多元化和数字化的发展道路，目前在国内外设有6个制造工厂，销售区域覆盖全球18个国家和地区，海外办事处主要有日本、新加坡、越南、印尼和德国等。

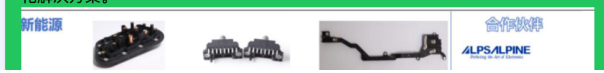
兴瑞电子的主要产品是什么

兴瑞电子的主要产品涉及多个领域，包括：

汽车电装领域：公司的汽车电子领域产品主要包括连接器、液晶框架、IBMU/BDU/DCDC等镶嵌注塑零组件、电控周边零组件及传感器组件等。这些产品服务于包括Panasonic、HELLA、BOSCH、TADITEL等国际汽车电子龙头公司，以及中车时代等国内Tier 1汽车电子客户。

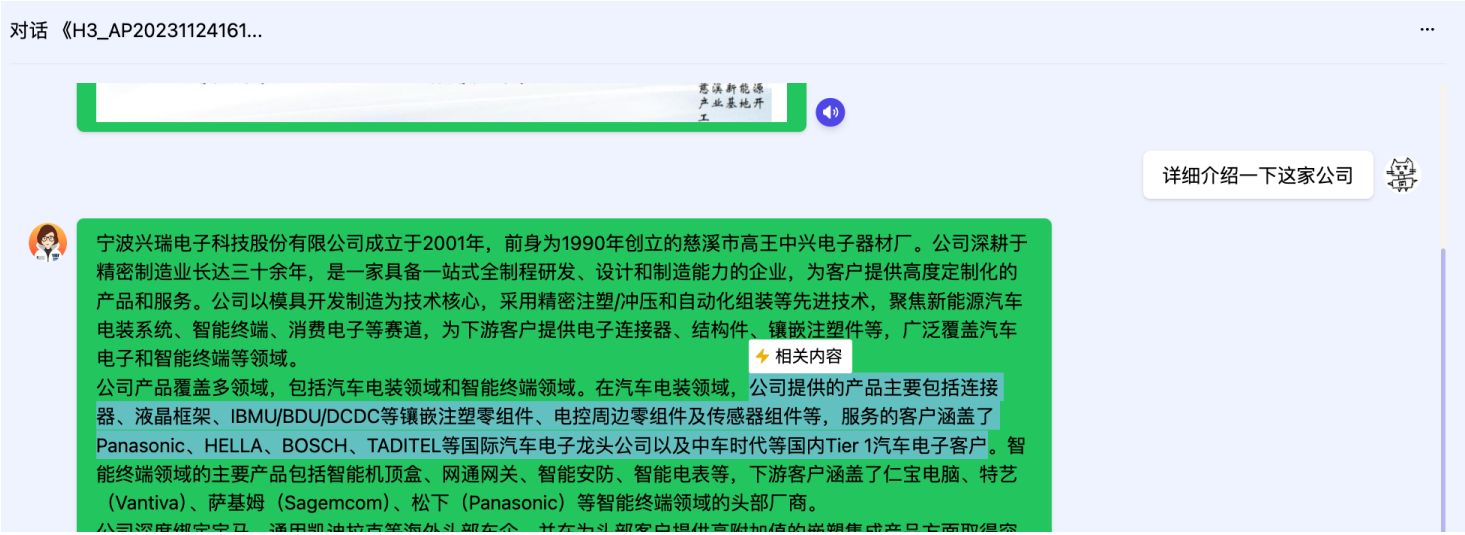
智能终端领域：在智能终端领域，公司的产品主要包括智能机顶盒、网通网关、智能安防、智能电表等品类。下游客户包括仁宝电脑、特艺（Vantiva）、萨基姆（Sagemcom）、松下（Panasonic）等智能终端领域的头部厂商。

公司以模具开发制造为技术核心，采用精密注塑/冲压和自动化组装等先进技术，为下游客户提供高度定制化的产品和服务，实现产品从零组件向模组的转变，并为行业高端客户提供定制化解决方案。



跨模态能力

集成视觉、听觉和其他感知模式的处理能力，以实现更丰富的用户体验和应用场景。



可解释的AI

开发技术来解释模型的决策过程，增加用户对模型输出的信任。

用RAG解决诸多问题

事实验证和误差校正

自然语言理解（NLU）的提升

多轮对话管理

动态知识库集成

在线学习

Serverless: Let's go beyond our capabilities

云端和边缘计算

模块化设计

Finetune: Stay hungry. Stay foolish.

域特定模型

逻辑推理

常识推理

算法与优化

- 算法和硬件协同优化
- 稀疏模型和压缩技术
- 优化算法和架构

商业合作

生态系统合作

开放API和平台

SEC案例

多语言支持

跨文化理解

Name and Principal Position	Year	Salary ($)	Bonus ($)	Stock Awards[1] ($)	Non-Equity Incentive Plan Compensation[2] ($)	All Other Compensation ($)	Total ($)
Tim Cook Chief Executive Officer	2023	3,000,000		46,970,283	10,713,450	2,526,112[3]	63,209,845
	2022	3,000,000	—	82,994,164	12,000,000	1,425,933	99,420,097
	2021	3,000,000	—	82,347,8			98,734,394
Luca Maestri Senior Vice President, Chief Financial Officer	2023	1,000,000		22,323,6		4)	26,935,883
	2022	1,000,000	—	22,132,0			27,151,798
	2021	1,000,000	—	21,959,6			26,978,503
Kate Adams Senior Vice President, General Counsel and Secretary	2023	1,000,000		22,323,6		5)	26,941,705
	2022	1,000,000	—	22,132,0			27,147,223
	2021	1,000,000	—	21,959,6			26,974,153
Deirdre O'Brien Senior Vice President, Retail	2023	1,000,000		22,323,6		6)	26,937,010
	2022	1,000,000	—	22,132,0			27,151,798
	2021	1,000,000	—	21,959,6			27,020,811
Jeff Williams Chief Operating Officer	2023	1,000,000		22,323,641	3,571,150	66,435[7]	26,961,226
	2022	1,000,000	—	22,132,015	4,000,000	18,337	27,150,352
	2021	1,000,000	—	21,959,620	4,000,000	17,437	26,977,057

td 10.96 × 21.33
Color #000000
Font 12px Arial, Helvetica, sans-serif
Padding 2.66667px 0px
ACCESSIBILITY
Contrast Aa 21
Name
Role gridcell
Keyboard-focusable

SEC案例

错误的格式需要被重新解析

SEC案例

SEC案例

Give me some tables on iPhone sales in 2019

	2019	Change	2018	Change	2017
Net sales by category:					
iPhone (1)	$ 142,381	(14)%	$ 164,888	18 %	$ 139,337
Mac (1)	25,740	2 %	25,198	(1)%	25,569
iPad (1)	21,280	16 %	18,380	(2)%	18,802
Wearables, Home and Accessories (1)(2)	24,482	41 %	17,381	36 %	12,826
Services (3)	46,291	16 %	39,748	22 %	32,700
Total net sales	$ 260,174	(2)%	$ 265,595	16 %	$ 229,234

	2019	2018	2017
iPhone (1)	$ 142,381	$ 164,888	$ 139,337
Mac (1)	25,740	25,198	25,569
iPad (1)	21,280	18,380	18,802
Wearables, Home and Accessories (1)(2)	24,482	17,381	12,826
Services (3)	46,291	39,748	32,700
Total net sales (4)	$ 260,174	$ 265,595	$ 229,234

给我一些关于iPhone销售的表格

	2019	Change	2018	Change	2017
Net sales by category:					
iPhone (1)	$ 142,381	(14)%	$ 164,888	18 %	$ 139,337
Mac (1)	25,740	2 %	25,198	(1)%	25,569
iPad (1)	21,280	16 %	18,380	(2)%	18,802
Wearables, Home and Accessories (1)(2)	24,482	41 %	17,381	36 %	12,826
Services (3)	46,291	16 %	39,748	22 %	32,700
Total net sales	$ 260,174	(2)%	$ 265,595	16 %	$ 229,234

	2019	2018	2017
iPhone (1)	$ 142,381	$ 164,888	$ 139,337
Mac (1)	25,740	25,198	25,569
iPad (1)	21,280	18,380	18,802
Wearables, Home and Accessories (1)(2)	24,482	17,381	12,826
Services (3)	46,291	39,748	32,700
Total net sales (4)	$ 260,174	$ 265,595	$ 229,234

知未智能出海

使用LLM建设新“巴别塔”

中书 · 技术

文档智能 & LLM & RAG

OpenRAG
https://medium.com/@yufan1602/modular-rag-and-rag-flow-part-%E2%85%B0-e69b32dc13a3

RAG

Retrieval-Augmented Generation for Large Language Models: A Survey

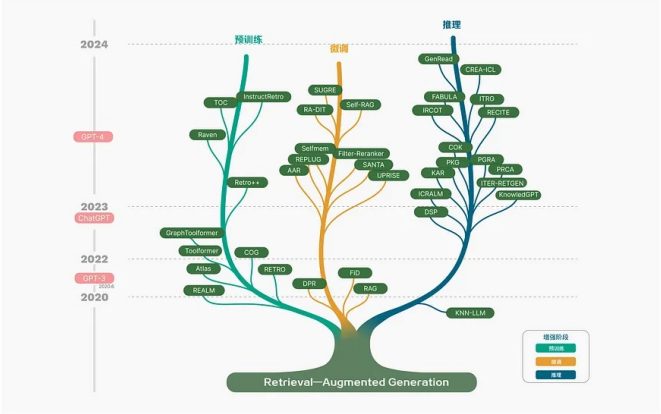


Figure 1: Technology tree of RAG research development featuring representative works

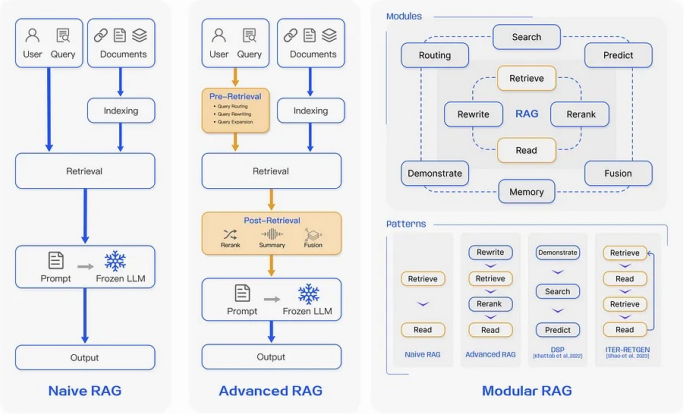


Figure 2: Comparison between the three paradigms of RAG

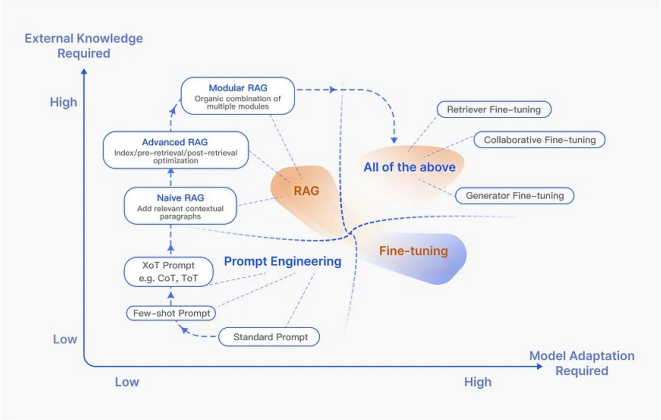


Figure 3: RAG compared with other model optimization methods

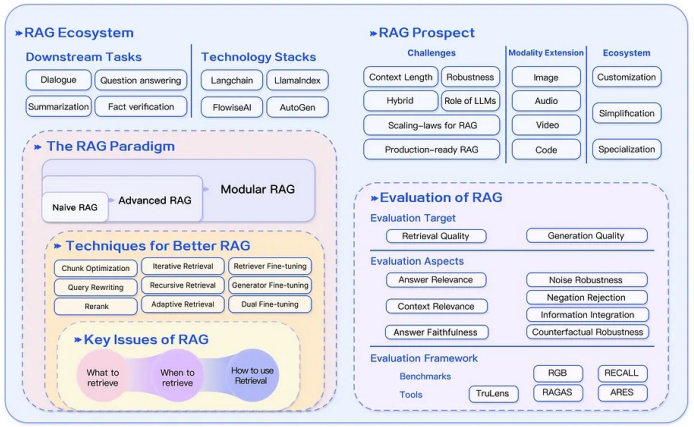
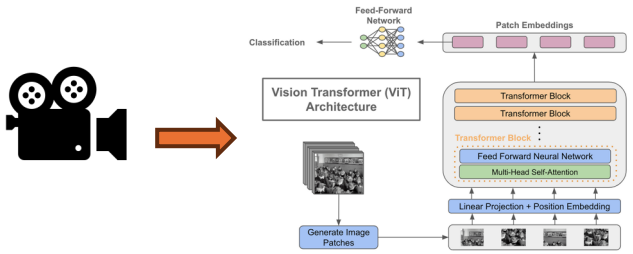


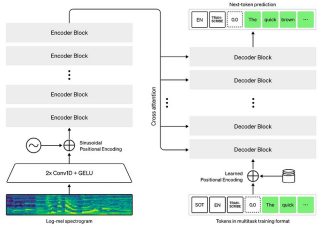
Figure 4: Summary of RAG ecosystem

多模态智能

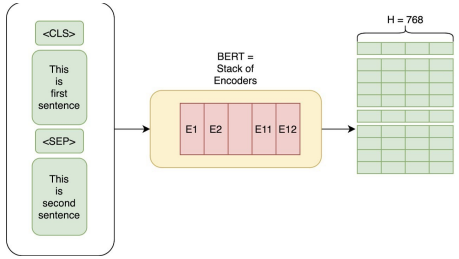


图像向量模型

0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

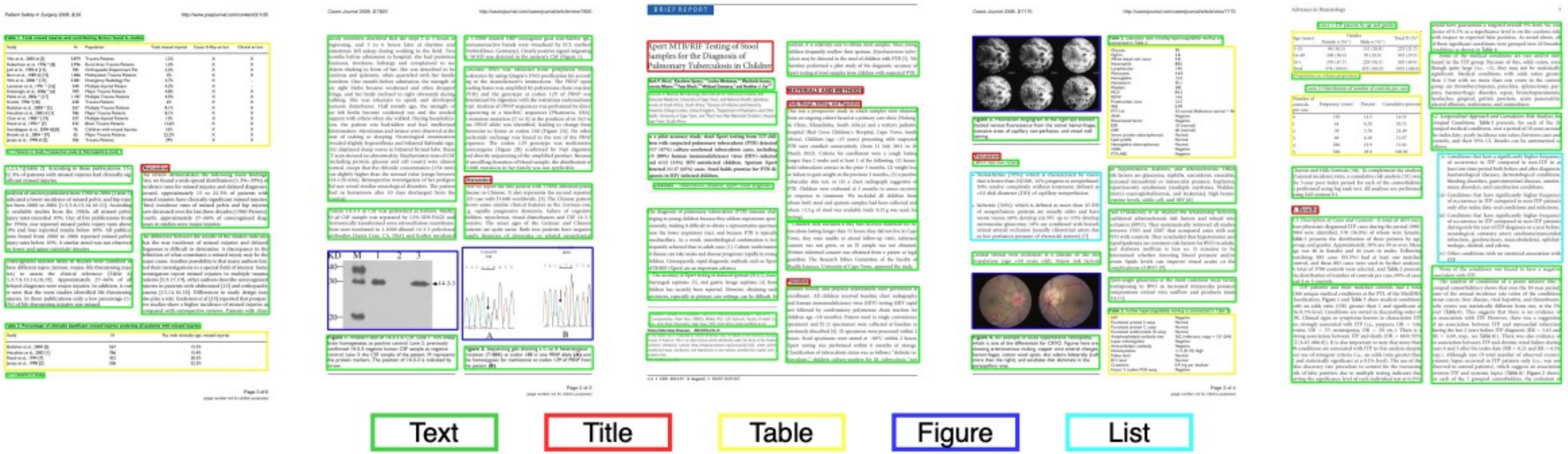


声音向量模型

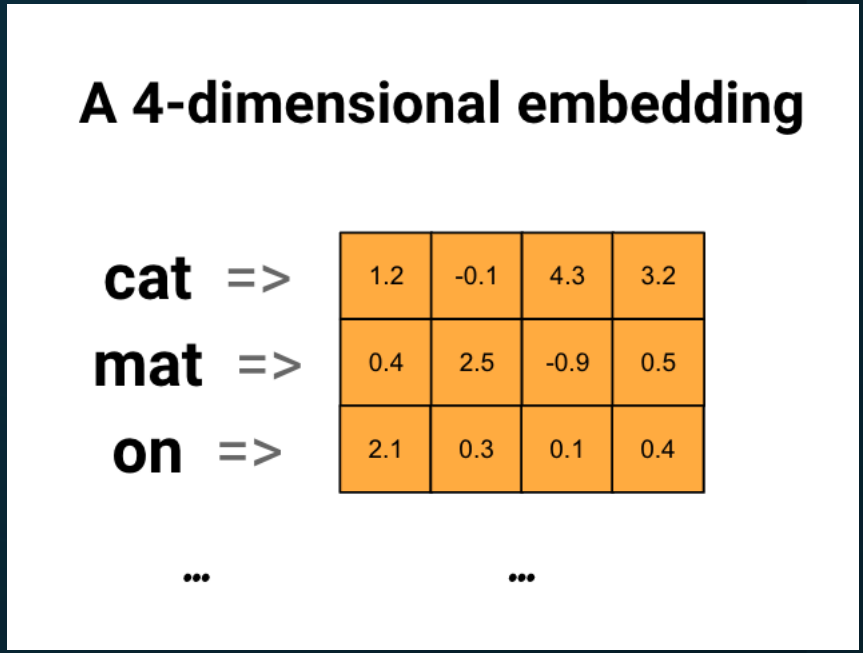


文本向量模型

多模态智能



向量搜索

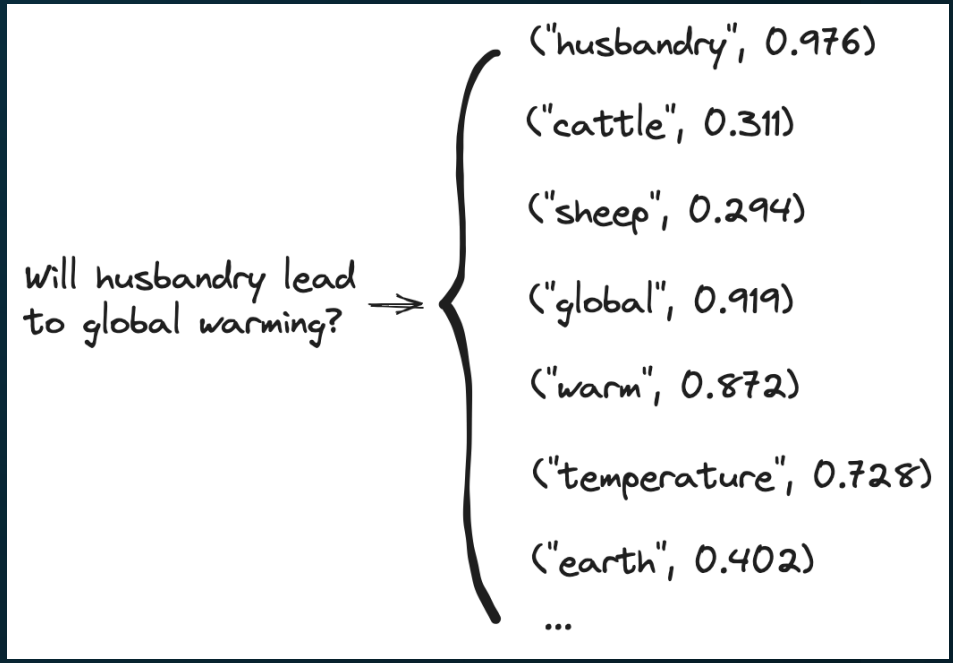

A 4-dimensional embedding
cat =>
mat =>
on =>
1.2 -0.1 4.3 3.2
0.4 2.5 -0.9 0.5
2.1 0.3 0.1 0.4
...
...

现实世界

稠密向量的分辨率不够

000001
000002
000003
000004
……

稀疏向量搜索


Will husbandry lead to global warming?
("husbandry", 0.976)
("cattle", 0.311)
("sheep", 0.294)
("global", 0.919)
("warm", 0.872)
("temperature", 0.728)
("earth", 0.402)
...

稀疏向量不同于传统字符搜索与倒排索引

- 我 / 爱 / 上海
- 上海 / 爱 / 我

全词模型：“600519的新股发行计划”

开源稀疏模型：

```
{
  “60”: 0.2,
  “05”: 0.1,
  “1”: 0.3,
  “9”: 0.4,
  “新股”: 0.6,
  “发行计划”: 0.4,
}
```

中书稀疏模型

```
{
  “600519”: 0.6,
  “新股”: 0.6,
  “发行”: 0.3,
  “发行计划”: 0.3,
  “新股发行”: 0.7,
  “新股发行计划”: 0.7,
}
```

稀疏向量搜索

23/06
Elasticsearch
支持

23/10
OpenSearch
支持

24/02
PGVecto支持

24/02 BGE开源同时支持稠密向量与稀疏向量模型

24/03
OpenSearch
开源稀疏向量模型

24/03 Milvus
2.4开始支持

24/04 中书:全词稀疏语义模型

模块化设计&Serverless&

SERVERLESS

混合SERVERLESS

LLM & SERVERLESS

Multi-Agent 历史进程

90% 👨 + 10% 🤖

让AI配合人

人类用大量时间引导AI
AI提供辅助资料

50% 👨 + 50% 🤖

让AI与人协作

AI与人共同建设

10% 👨 + 90% 🤖

让AI完成大部分工作

AI完成绝大部分工作

Multi-Agent 历史进程

100% 👨 + 11% 🤖

让AI配合人

1.1倍生产力

100% 👨 + 100% 🤖

让AI与人协作

2倍生产力

100% 👨 + 900% 🤖

让AI完成大部分工作

10倍生产力

谢谢